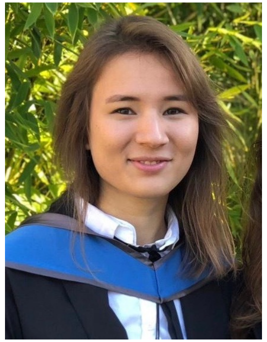
Speedy Performance Estimation for Neural Architecture Search

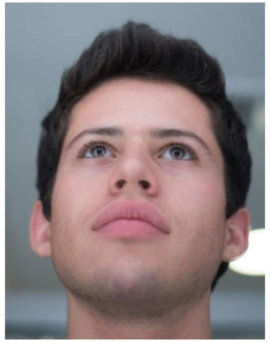

Binxin Ru*


Clare Lyle*


Lisa Schut

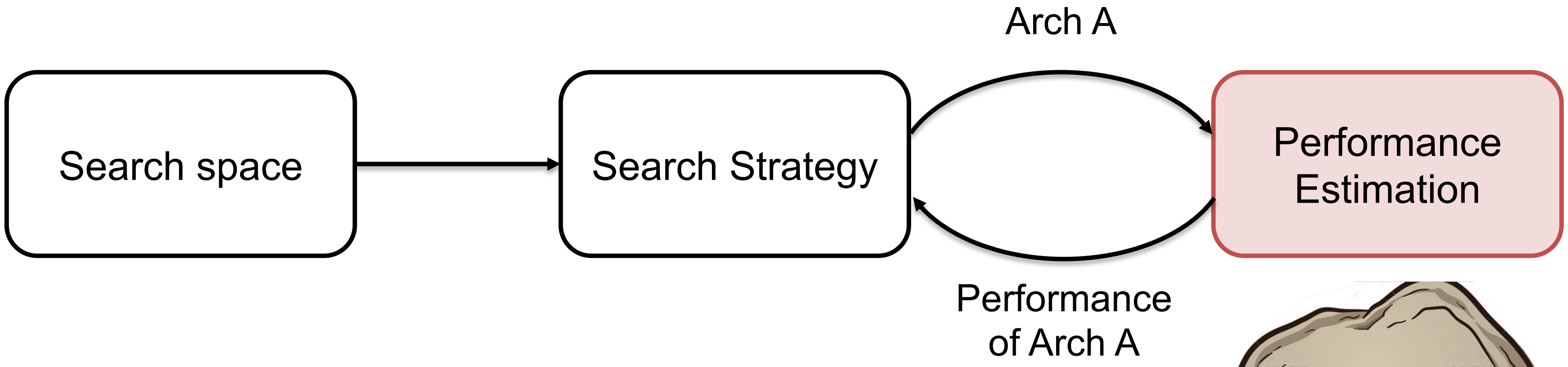

Miroslav Fil

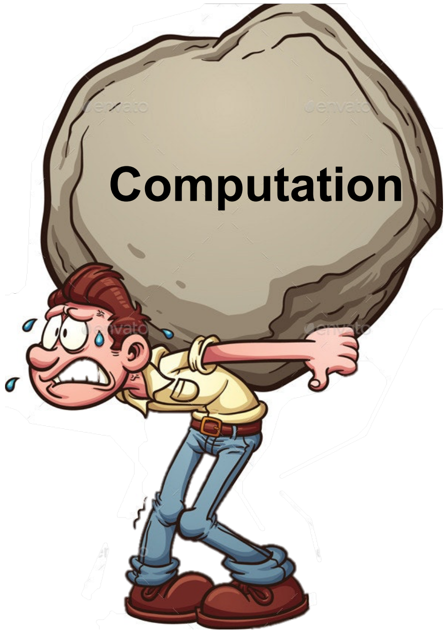

Mark van der Wilk


Yarin Gal

Computation bottleneck of NAS

- *<u>Reliable</u>* yet *<u>efficient</u>* estimation of generalisation performance of a proposed architecture

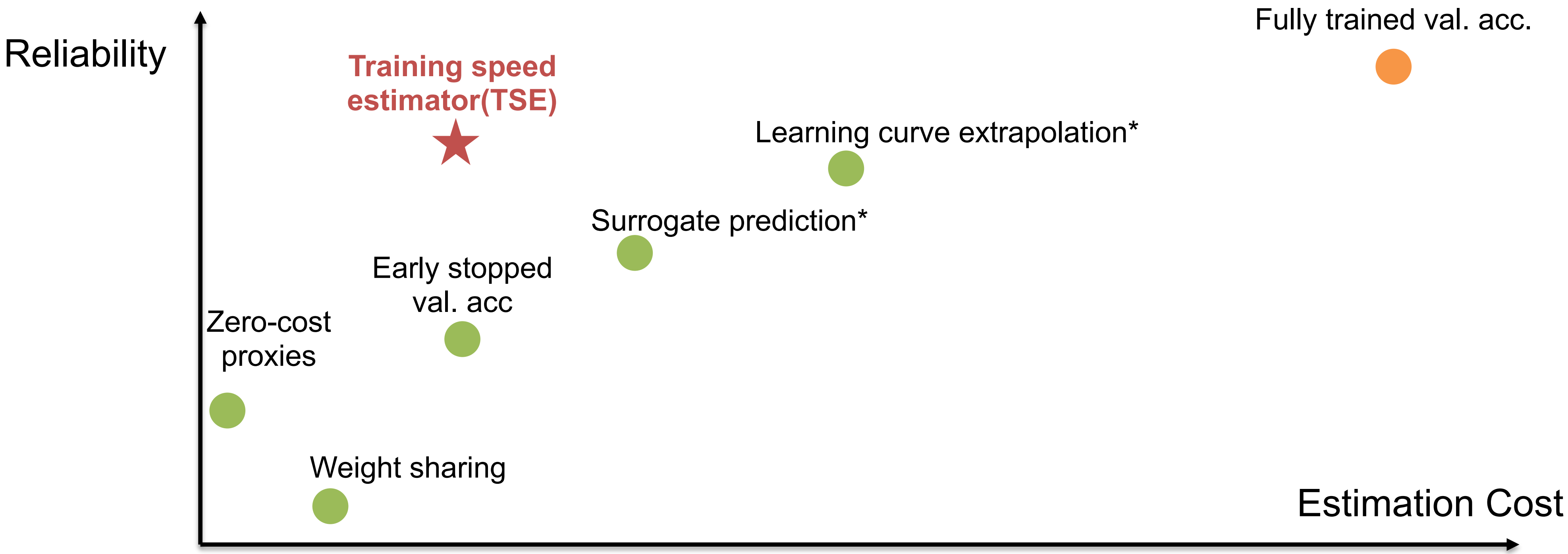




(Elsken et al., 2019)

Performance Estimation Methods

- **Reliability**: how well the *estimated* performance correlates with the *true test* performance
- **Costs**: costs for computing the estimates and/or collecting architecture data for surrogate training/tuning
- A *simple, cheap, theoretically-motivated, reliable* solution: Training speed estimator (TSE)



(Zoph et al., 2017; Real et al., 2017; Pham et al., 2018; Liu et al., 2019; Baker et al., 2017; Mellor et al., 2020; Abdelfattah et al., 2021)

Training Speed Estimator (TSE)

- The generalisation performance of an architecture can be estimated via a simple measure of training speed, the sum of its SGD training losses over first T epochs

$$\mathrm{TSE} = \sum_{t=1}^{T} \left[\frac{1}{B} \sum_{i=1}^{B} \ell \left(f_{\theta_{t,i}}(\mathbf{X}_i), \mathbf{y}_i \right) \right]$$

- Two variants of TSE to account for unstable dynamics in very early training:

$$\mathrm{TSE\text{-}E} = \sum_{t=T-E+1}^{T} \left[\frac{1}{B} \sum_{i=1}^{B} \ell \left(f_{\theta_{t,i}}(\mathbf{X}_i), \mathbf{y}_i \right) \right]$$

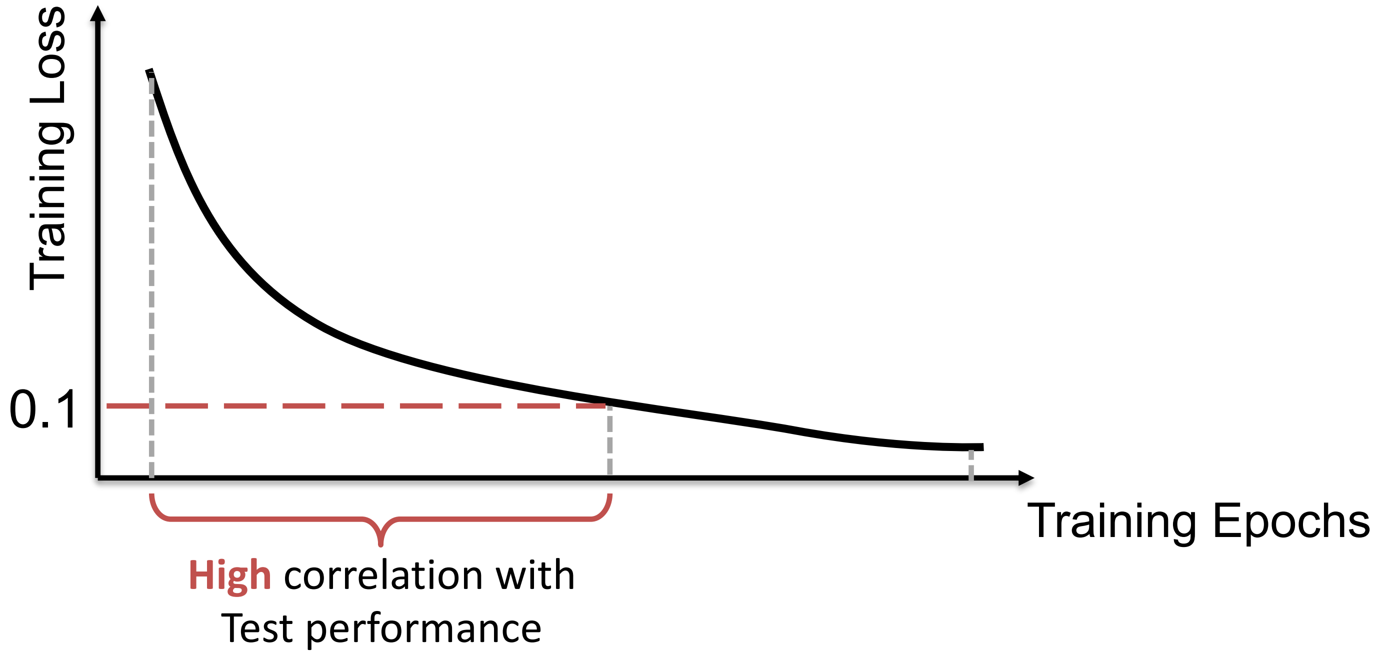
Only consider the most recent E epochs

$$\mathrm{TSE\text{-}EMA} = \sum_{t=1}^{T} \gamma^{T-t} \left[\frac{1}{B} \sum_{i=1}^{B} \ell \left(f_{\theta_{t,i}}(\mathbf{X}_i), \mathbf{y}_i \right) \right]$$

Downweigh earlier training trajectory

Theoretical Motivations

- Training Speed and Generalisation
  - Train faster (fewer optimization steps) ⇔ generalize better (Hardt et al., 2016)
  - No. of steps to reach certain training loss ⇔ test performance (Jiang et al., 2020)



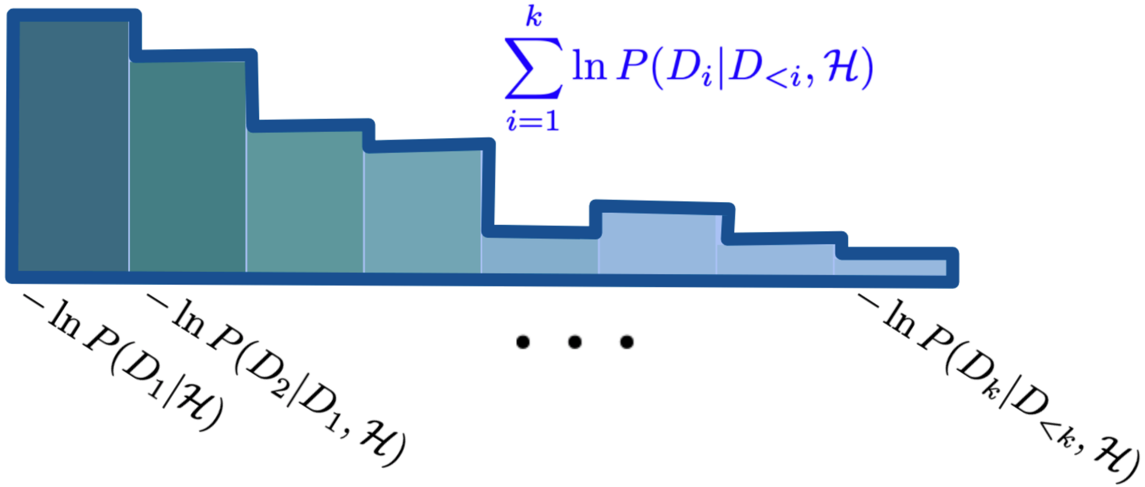
(Hardt et al., 2016; Jiang et al., 2021)

Theoretical Motivations

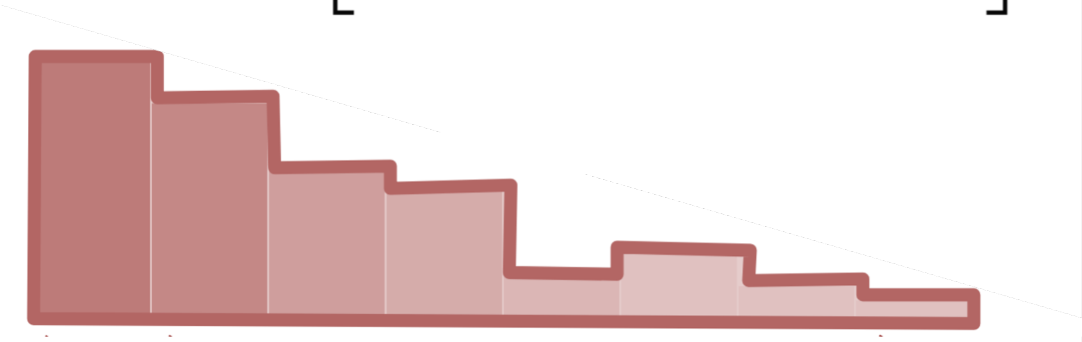
- Training Speed and Generalisation
- Bayesian Marginal Likelihood of a model $\mathcal{H}$
  - No need for validation data

$$\log P(\mathcal{D}|\mathcal{H}) = \sum_{i=1}^{n} \overbrace{\log P(\mathcal{D}_i|\mathcal{D}_{<i}, \mathcal{H})}^{\text{sum of log likelihoods}}$$

$$\text{TSE} = \sum_{t=1}^{T} \left[\frac{1}{B} \sum_{i=1}^{B} \overbrace{\ell\left(f_{\theta_{t,i}}(\mathbf{X}_i), \mathbf{y}_i\right)}^{\text{sum of training losses}} \right]$$



Bayesian update steps

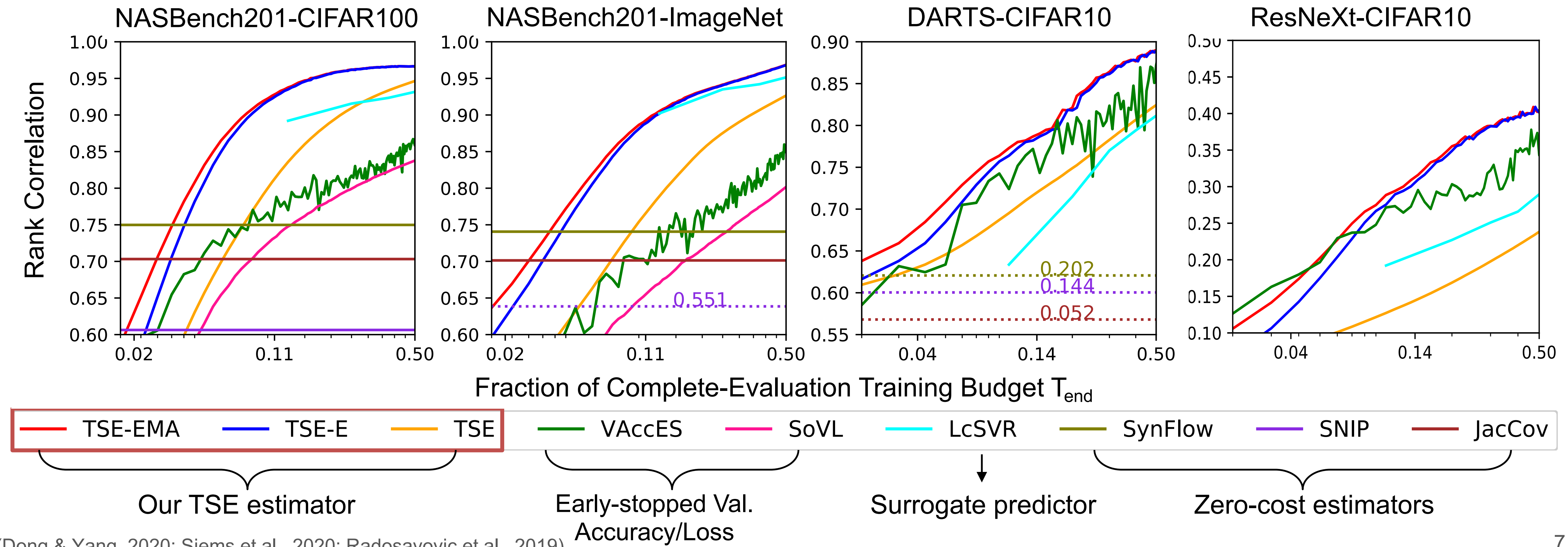


SGD training steps

(Lyle et al., 2020)

Exp1: Rank Correlation with Generalization

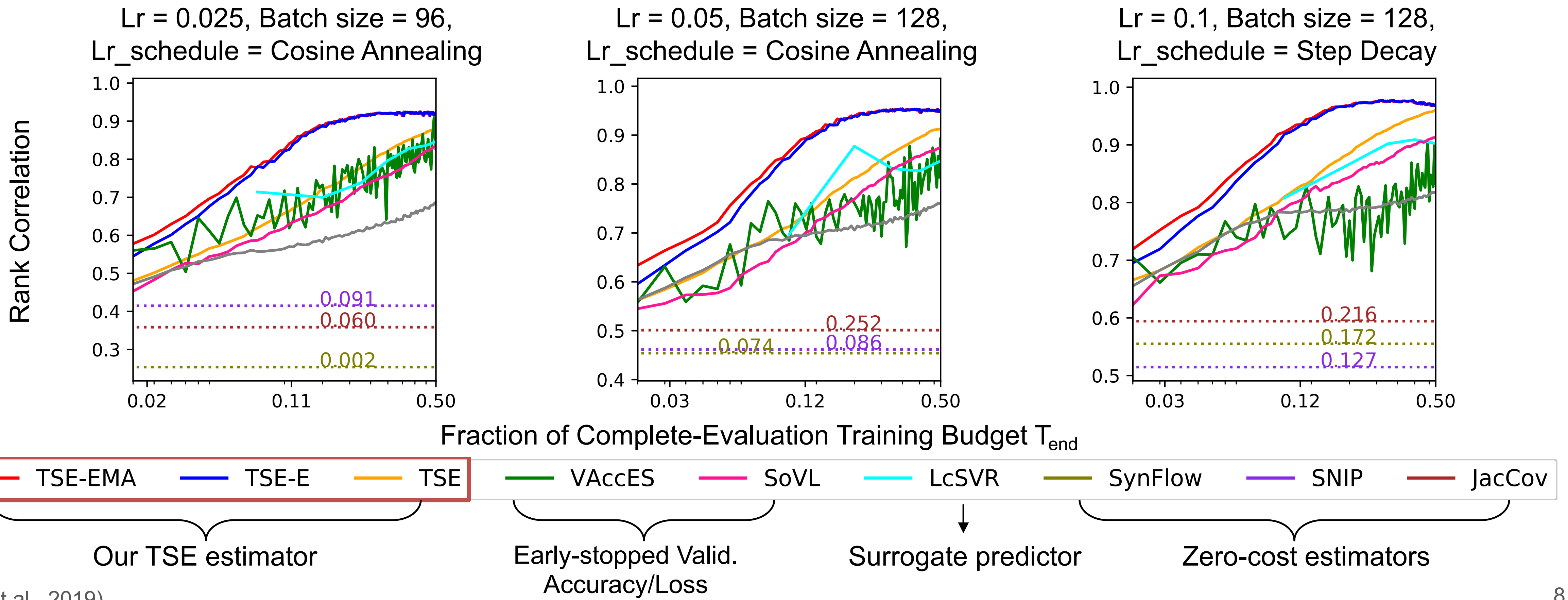
- Rank correlation: between estimated *ranking* and true test accuracy *ranking*
- Consistently outperform all competing estimators on a diverse set of search spaces and image tasks, with a small fraction of training budget



(Dong & Yang, 2020; Siems et al., 2020; Radosavovic et al., 2019)

Exp1: Rank Correlation with Generalization

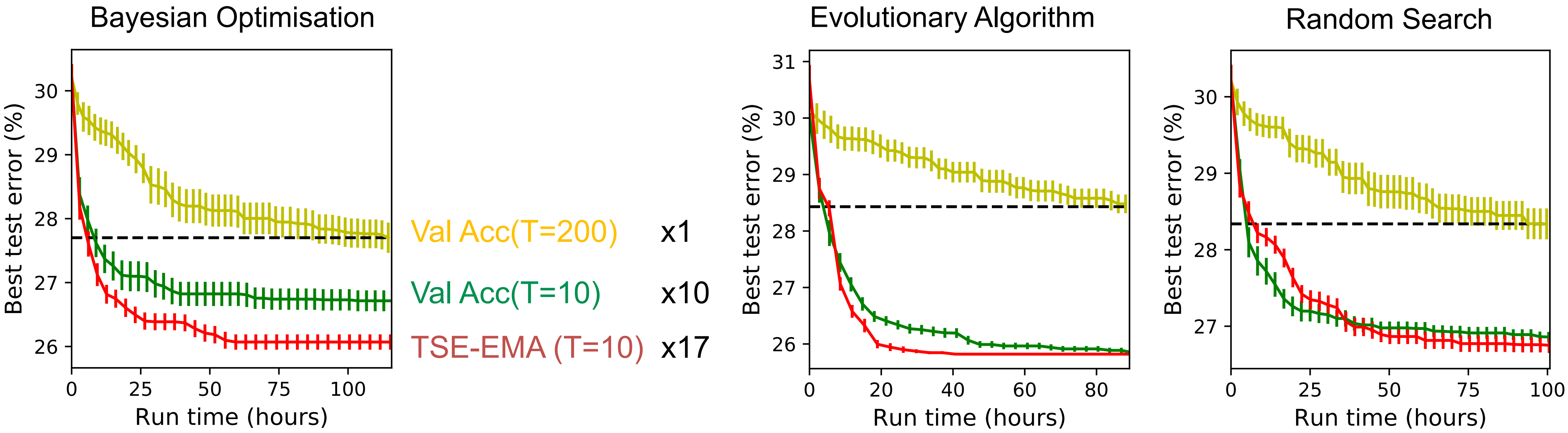
- *Rank* correlation between estimator and true test accuracy
- Robust to various training set-ups on DARTS search space
  - Optimal architectures under one training protocol may not remain optimal in another



(Liu et al., 2019)

Exp2: Improve Query-based NAS

- Evaluate architecture performance at each query using TSE-EMA (T=10) vs early-stopped val. Accuracy Val Acc(T=10) and fully-trained val. accuracy Val Acc(T=200)
- Significantly reduce search costs for different query-based search strategies



No exploitation

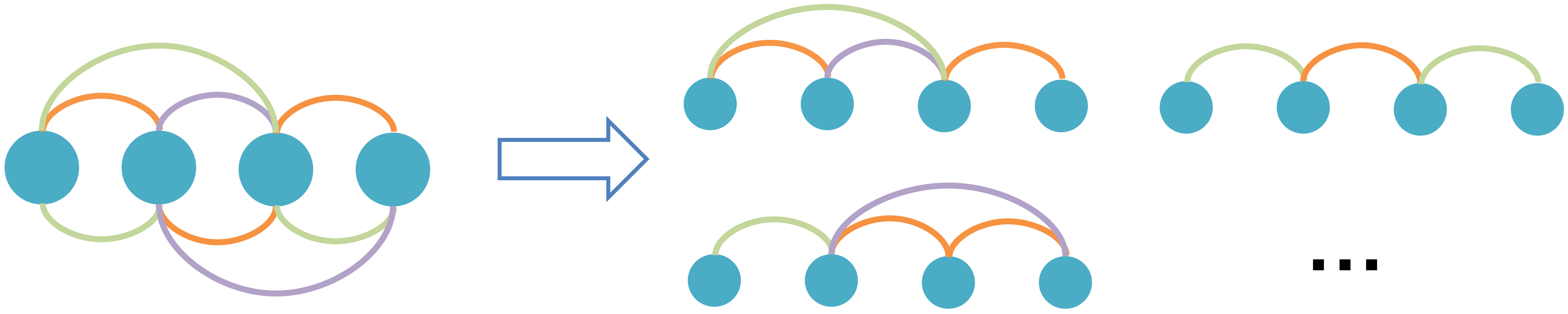
(Ying et al., 2019)

Exp2: Improve One-shot NAS

- *Replace* validation accuracy with TSE score for subnetwork evaluation

Phase 1
Super-Network Training

Phase 2
Sub-Network(Architecture) Search

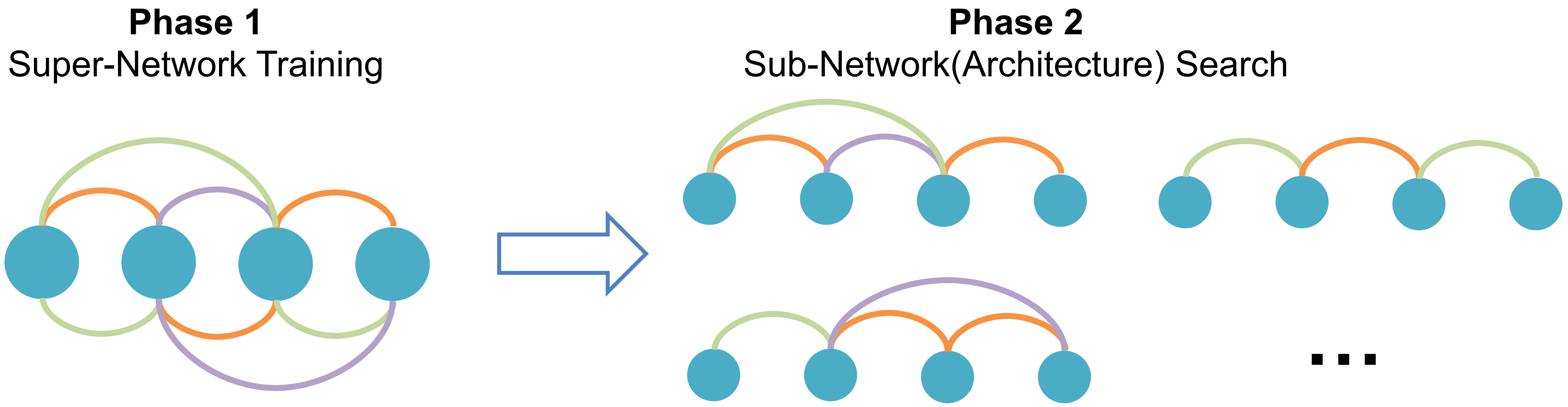


Evaluate each subnetwork by inherit *corresponding* super-net weights
→ compute valid. accuracy

(Shi et al., 2020)

Exp2: Improve One-shot NAS

- *Replace* validation accuracy with TSE score for subnetwork evaluation



Evaluate each subnetwork by inherit *corresponding* super-net weights
→ ~~compute valid. accuracy~~
→ compute TSE by train for B additional mini-batches*

* Remove the cost and need for running on validation data

(Shi et al., 2020)

Exp2: Improve One-shot NAS

- Replace valid. accuracy with TSE score for subnetwork evaluation
  → compute TSE by train for B additional mini-batches
- Better estimate of the subnetworks' true ranking* & Lead to better top architectures on both NASBench201(*NB201*) and NASBench301(*DARTS*)
- Orthogonal to super-net training techniques (*RandNAS, FairNAS, MultiPaths*)

B	Estimator	Rank Correlation				Average Accuracy of Top 10 Architectures			
		NB201-CIFAR10			DARTS	NB201-CIFAR10			DARTS
		RandNAS	FairNAS	MultiPaths	RandNAS	RandNAS	FairNAS	MultiPaths	RandNAS
100	TSE	**0.70 (0.02)**	**0.84 (0.01)**	**0.83 (0.01)**	**0.30(0.04)**	**92.67 (0.12**	**92.7 (0.1)**	**92.63 (0.12)**	**93.64(0.04)**
	Val Acc	0.44 (0.15)	0.56 (0.17)	0.67 (0.05)	0.11(0.04)	91.47 (0.31)	91.73 (0.21)	91.77 (0.78)	93.20(0.04)
200	TSE	**0.70 (0.03)**	**0.850 (0.01)**	**0.83 (0.01)**	**0.32(0.04)**	**92.70 (0.00)**	**92.77 (0.06)**	**92.73 (0.06)**	**93.55(0.04)**
	Val Acc	0.41 (0.10)	0.56 (0.17)	0.53 (0.11)	0.09(0.02)	91.53 (0.55)	92.40 (0.10)	92.23 (0.23)	93.34(0.02)
300	TSE	**0.71 (0.03)**	**0.851 (0.00)**	**0.82 (0.01)**	**0.34(0.04)**	**92.70 (0.00)**	**92.77 (0.06)**	**92.70 (0.00)**	**93.65(0.04)**
	Val Acc	0.44 (0.04)	0.62 (0.08)	0.59 (0.71)	0.06(0.02)	91.20 (0.35)	92.10 (0.50)	91.43 (0.72)	93.31(0.02)

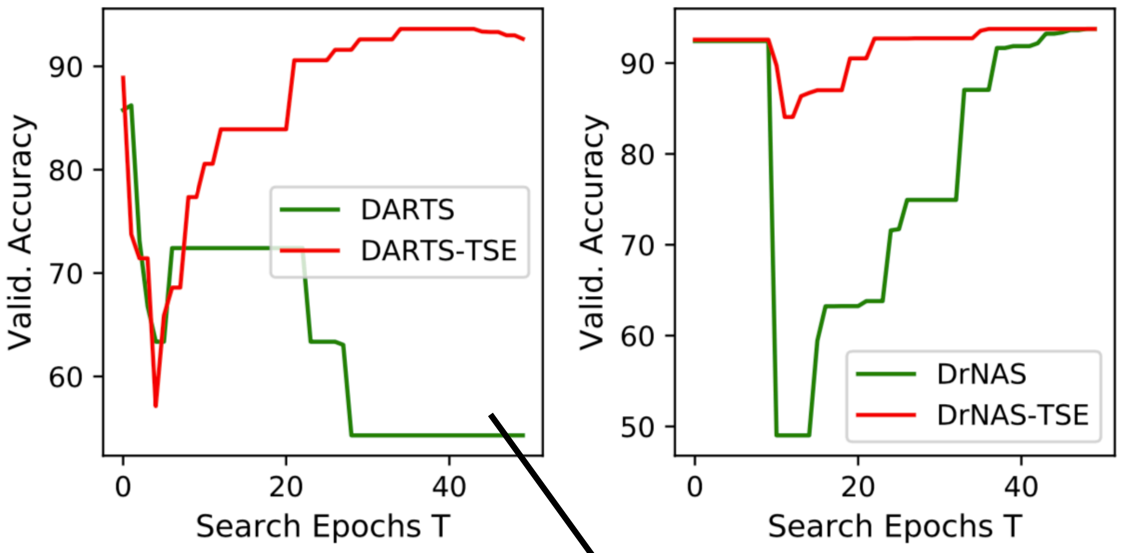
* When training the subnetwork from scratch

(Li et al., 2020; Chu et al., 2019; Yu et al., 2019)

Exp2: Improve Differentiable NAS

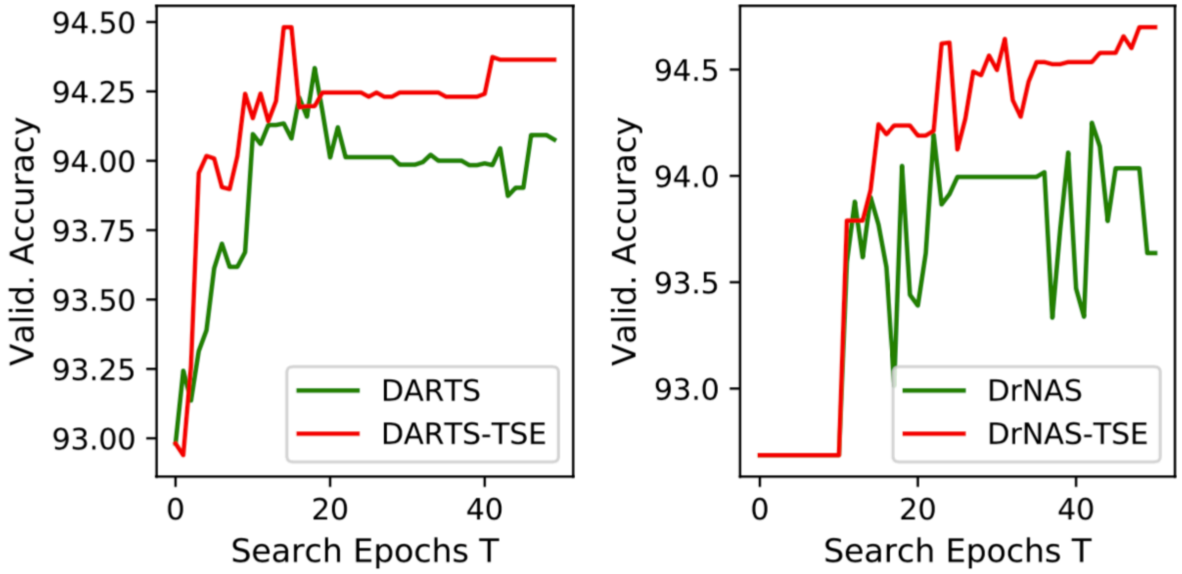
- Use the gradient of ~~validation loss~~ TSE to update architecture parameters
  - → Improve search performance & Mitigate overfitting to *skip-connect* operation
- Modified DARTS(left) and DrNAS (right)

NASBench201-CIFAR10



All operations become *skip connect*

NASBench301-CIFAR10



(Liu et al., 2019; Chen et al., 2021)

Summary

- We propose a novel performance estimator, **TSE** :

Simple & Cheap Easy to compute, Require very small amount of training, Model-free!	**Theoretically-motivated** Training speed, Bayesian marginal likelihood
Reliable & Robust Good rank correlation on various search spaces, image tasks, training set-ups	**General** Easily applicable to various NAS approaches, including query-based, one-shot, differentiable

- Paper link: https://arxiv.org/abs/2006.04492
- Code: https://github.com/rubinxin/TSE